Chapter 8
Assessing Relationships with Correlation and Regression

Section 8.2
The Regression Model

David W. Gerbing

The School of Business
Portland State University

- Regression Analysis
  - The Model
  - Estimation of the Model
  - Inferential Analysis

8.2a
The Model

Application of Regression Analysis

Many uses regression analysis as a form of machine learning

- Two primary goals of regression analysis, usually of which one is the primary focus of a given analysis: Build a model to
  - Forecast the unknown value of response variable y from one or more predictor variables X (capital X because can be, and usually is, more than 1 variable)
  - Explain why the value of response variable y is obtained in terms of the relations among the predictor variables X to each other and to the y
- Regression analysis has many, many applications in management and economics, including most instances of forecasting . . .
  - demand for more hospital beds as population size increases
  - demand for inventory items at different times of the year
  - selling price of a house based on size and age

Regression Compared to Correlation

Two complementary procedures

- Both regression and correlation provide evidence regarding the strength of a relationship between variables
- Size of correlation coefficient indicates extent of linear relationship, shown by the width of a confidence ellipse
- **Regression analysis**: Estimate a linear equation or model, plotted as the best line through the scatterplot of X and y
- Regression analysis indicates the extent of a relationship by how accurately the model (i.e., line) estimates the value of y from the corresponding value of X
- **Multiple regression analysis**: Estimate and analyzes a model of y from *two* or more X's
- **Key Concept**: Prediction complements correlation, but neither prediction nor correlation imply causation

Unconditional and Conditional Means

Analyze Distribution of y separately for each value of X

- **Unconditional Mean**: The mean of all the data, m
- The unconditional mean is just the regular mean, but the term is introduced to set up the following concept
- **Conditional Mean**: Mean of y for just those data values with a specific value of X
- Consider relating Height and Weight
  - The mean Weight for all people in the sample is the unconditional mean of Weight
  - The mean Weight just for the people who are 68 inches tall is the conditional mean of Weight for $x_{Ht} = 68$

The Mean as the Basis of the Forecast

Use information about X to forecast *y*

- **Key Concept**: The forecast of a value of *y* is the mean of *y*
  - If there is no information regarding variable X, or if X is unrelated to y, the forecast of *y* is the *unconditional* mean
  - If X and *y* are related, the forecast is the *conditional* mean of *y* for that specific value of X from which to forecast the corresponding value of y
- If you had enough data at each value of X, *x* for a single variable as in this example, then calculate forecasts as simple averages

Illustration: Conditional Mean, Data

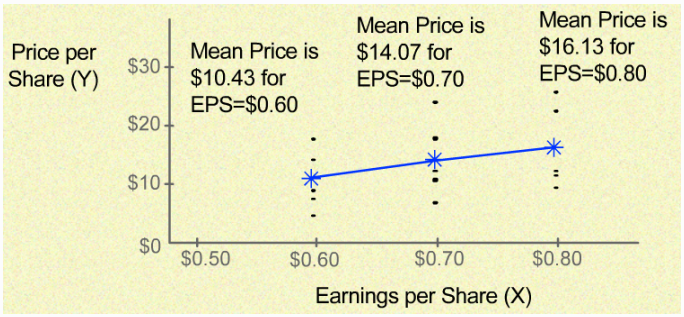
Consider Price per Share (PPS) and Earnings per Share (EPS)

- Analyze PPS for three levels of EPS: \$0.60, \$0.70, \$0.80

	Company	1992 EPS X	Closing 1992 PPS Y	Conditional Mean of Y
1	BRANDON SYSTEMS CORP	\$0.60	\$7.38	for companies with an EPS of \$0.60, $\bar{Y}_{.60}$ = \$10.43
2	BUSH INDUSTRIES	\$0.60	\$8.75	
3	TEXAS PACIFIC LAND TRUST	\$0.60	\$17.38	
4	BLOUNT INC	\$0.60	\$14.00	
5	PENOBSCOT SHOE	\$0.60	\$4.63	
6	CALGON CARBON CORP	\$0.70	\$17.63	for companies with an EPS of \$0.70, $\bar{Y}_{.70}$ = \$14.07
7	SAMSON ENERGY CO	\$0.70	\$10.25	
8	MEDIA GENERAL	\$0.70	\$17.38	
9	GENCORP INC	\$0.70	\$10.63	
10	SCIENTIFIC-ATLANTA INC	\$0.70	\$23.75	
11	BAIRNCO CORP	\$0.70	\$6.75	
12	DANIEL INDUSTRIES	\$0.70	\$12.13	
13	COLES MYER LTD	\$0.80	\$25.63	for companies with an EPS of \$0.80, $\bar{Y}_{.80}$ = \$16.13
14	MANITOWOC CO	\$0.80	\$22.25	
15	ROLLINS TRUCK LEASING	\$0.80	\$12.00	
16	CANADIAN MARCONI CO	\$0.80	\$11.38	
17	SPORT SUPPLY GROUP INC	\$0.80	\$9.38	

Illustration: Conditional Mean, Plot

Scatterplot with conditional means



- The conditional means, the forecasted values of PPS based on EPS, are almost linearly related

Conditional Means Plus Linearity Assumption

Apply concept of conditional means with an assumption

- Usually not enough data values for each value of X to estimate the corresponding conditional mean directly from the data
- Instead, assume a functional relationship which is usually linearity
- **Key Concept**: Every point on the regression line is a conditional mean
- Assuming linearity, each point on the line is a conditional mean that could hypothetically be estimated directly as the mean of y just for those rows of data that have a specified value of X

The Regression Model

Basic terminology and notation

- **Function**: Relationship among variables in which the value of one (or more) variables, X, exactly determines the value of another variable, y
- **Regression model**: From a sample of paired data values for variables X and y, estimate the coefficients of the linear function that calculates $\hat{y}$ given the value of X
  - General linear form for one X variable: $\hat{y} = b_0 + b_1 x$
  - A specific linear equation: $\hat{y} = 1.5 + 10x$
- From the data values of X and y, the regression analysis provides sample values that estimate the y-intercept and slope of the model
  - b_0, the y-intercept
  - b_1, the slope coefficient

The Regression Model

Basic terminology and notation

- **Predictor variable**: X, the data values entered into the equation, also called the independent variable, explanatory variable, or feature
- **Response variable**: y, the variable to be explained, also called the dependent variable, explained variable, outcome variable, or label
- **Fitted value**: A value of $\hat{y}_i$ calculated from the model, where $\hat{y}_i = b_0 + b_1 x_i$ for a model with one predictor variable
- **Slope coefficient**: Change in $\hat{y}$ for each unit increase in X, which is the *average* change in the data values of y
- Estimated from the data, the sample slope coefficient b_1 describes the average change of y *only* for the particular data set from which the model was estimated
- **Intercept**: b_0, value of $\hat{y}$ when $x = 0$

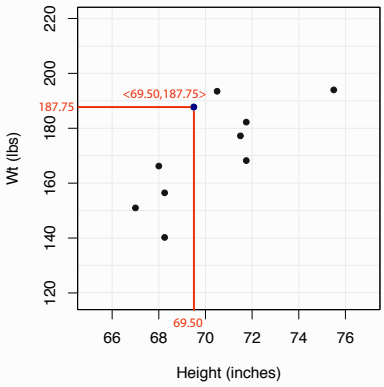
8.2b
Estimation of the Model

Ex: Data for Regression Analysis with Scatterplot

- Height (x) and Weight (y) for ten men:

	Ht	Wt
1	71.75	182.25
2	71.75	168.25
3	75.50	194.00
4	68.25	140.25
5	68.25	156.50
6	69.50	187.75
7	70.50	193.50
8	71.50	177.25
9	67.00	151.00
10	68.00	166.25

- Coordinates of one point, for the data for one person, here from row six of data, $< x_6, y_6 >$: The man with a Height of 69.50 inches and a Weight of 187.75 lbs



R: Regression of y on X with `lessR` `reg` function

R/lessR instructions for regression

- The `lessR` `Regression` function does regression analysis, here illustrated for response variable y and one predictor variable x

```
Regression(y ~ x)    or    reg(y ~ x)
```

- The ~ means "depends on" or "explained by", which indicates a model in R notation

```
reg.brief(y ~ x) for briefer output
```

- Following are the R/lessR instructions for the regression with briefer output for the 10 paired values of Height and Weight

```
library("lessR")
d <- Read("http://lessRstats.com/data/bodyfat10.csv")
reg.brief(Wt ~ Ht)
```

Regression Output: Estimated Model

Selectively edited output for the estimates of the model

```
Response Variable:   Wt
Predictor Variable:  Ht

             Estimate
(Intercept) -193.452
Ht             5.202
```

- The sample statistics, the values of intercept and slope estimated from the data: $b_0 = -193.45$ and $b_1 = 5.20$, so the estimated regression model is

  $$\hat{y} = -193.45 + 5.20x$$

- **Descriptive Result**: The *sample* slope coefficient, $b_1 = 5.20$, indicates that, for this *one sample of data*, an increase of Height of 1 inch yields an average increase of Weight of 5.2 lbs

Ex: Fitted Value from Regression Line

What value of Weight is consistent with a specific Height?

- Consider a value of the predictor variable, Height of 6^{th} man in the data set: 69.5 inches
- To get the fitted value of the response variable, Weight, enter $x_{HT} = 69.5$ into the model to calculate $\hat{y}_{WT}$,

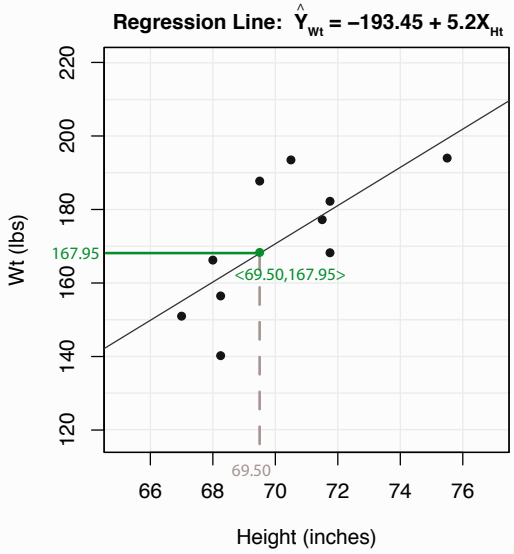
  $$\hat{y}_{WT} = -193.45 + 5.20x_{HT} = -193.45 + 5.20(69.5) = 167.95$$

- **Descriptive Result**: For the man in these data with a Height of 69.5 inches, the fitted value is a Weight of 167.95 lbs
- The coordinate of the 6^{th} point on the regression line,

  $$< x_6, \hat{y}_6 >=< 69.5, 167.95 >$$

- Note that in this context $\hat{y} = 167.95$ lbs is *not* a literal forecast of Weight for the 6^{th} man because we already know his *actual* weight, $y = 187.75$ lbs, from the data

Ex: Fitted Value from Regression Line



Distinction Between Actual and Fitted Values of y

Data in a scatterplot do not fall on a line

- A plot of the pairs of data points is a scatterplot from which the "best-fitting" line is calculated
- For each data value x_i of predictor variable x, there is
  - The data, a point in the scatterplot $< x_i, y_i >$
  - The corresponding value on the regression line $< x_i, \hat{y}_i >$
- **Modeling error** or **residual**: $e_i = y_i - \hat{y}_i$, difference between actual and fitted values of y for a given x_i
- One important goal of the analysis is to obtain a model that has small modeling errors

Ex: Error, the Distinction between y and $\hat{y}$

Illustrated on following slide

- Distinguish between what occurred for $x = 69.50$, the data,

  $y = 187.75$,

  and what is consistent with the underlying regression model, the fitted value,

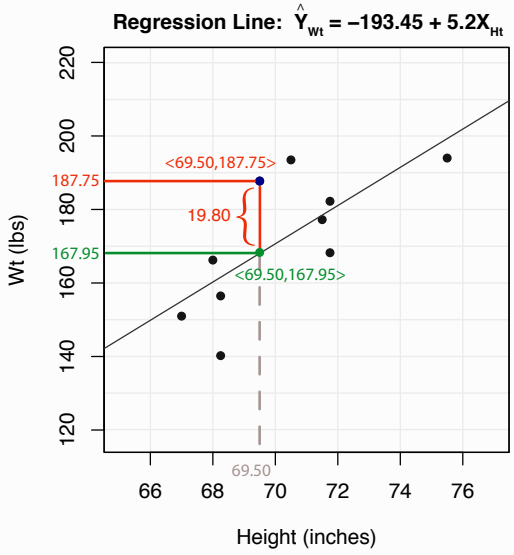
  $\hat{y} = 167.95$
- To get the modeling error, the residual, for the sixth row of data,

  $e_6 = y_6 - \hat{y}_6$,

  $= 187.75 - 167.95$,

  so $e_6 = 19.80$ lbs, the amount of underestimate by the model

Ex: Error, the Distinction between y and $\hat{y}$



Criterion Used to Construct a Regression Line

Estimate b_0 and b_1 in the equation $\hat{y} = b_0 + b_1 x$

- For each observation, a paired value of X and y, calculate $\hat{y}$, and then the corresponding residual term, the error $e = \hat{y}$
- For example, 10 rows of data yields 10 values of e
- $\sum e_i^2$ is the sum of squared (modeling) errors or SSE
- **OLS**: Ordinary Least Squares, the usual choice, from among many, for estimating the coefficients of the regression model
- The OLS criterion chooses the sample coefficients b_0 and b_1 that provide the minimal possible value of the sum of the squared residuals for that specific sample
  estimate model that minimizes: $\sum(y_i - \hat{y}_i)^2 = \sum e_i^2$
- The resulting estimated model that uses the OLS estimates, b_0 and b_1, defines the "best-fitting line" through the scatterplot
- Any other choice of values for b_0 and b_1 would result in *more* cumulative squared modeling errors, a larger value of $\sum e_i^2$

Sum of Squared Errors for $\hat{y} = -193.45 + 5.20x$

Illustration of the meaning of the sum of squared errors

- Given a value for b_0 and for b_1, can compute the sum of squared residuals, Σe^2

Ht	*Wt*	*PredWt*	*Residual*	*Residual^2*	
71.75	182.25	179.76248	2.488	6.188	
71.75	168.25	179.76248	-11.512	132.537	
75.50	194.00	199.26847	-5.268	27.757	
68.25	140.25	161.55688	-21.307	453.983	
68.25	156.50	161.55688	-5.057	25.572	
69.50	187.75	168.05888	19.691	387.740	
70.50	193.50	173.26048	20.240	409.638	
71.50	177.25	178.46208	-1.212	1.469	
67.00	151.00	155.05488	-4.055	16.442	
68.00	166.25	160.25648	5.994	35.922	
			0.000	1497.249	Sum

- Sum of squared residuals is 1497.25
- Any other choice of $b_0 = -193.45$ and $b_1 = 5.20$ results in a larger sum of squared residuals for these data

8.2c
Inferential Analysis

Inferential Considerations

What happens in the population?

- There is a question *always* to be asked of any descriptive statistic, such as b_0 or b_1
- How close is the calculated descriptive statistic likely to be to the unknown, but desired, population value?
- **Key Concept**: If a new sample of X and y values were taken, a *different* value of the slope coefficient b_1 would be obtained
- Like any descriptive statistic, the slope coefficient b_1 has a standard error, the standard deviation of the statistic across repeated samples, here s_{b_1}
- The basis of statistical inference is to understand the sampling variability of the descriptive statistic, that is, its standard error

Population Regression Coefficients

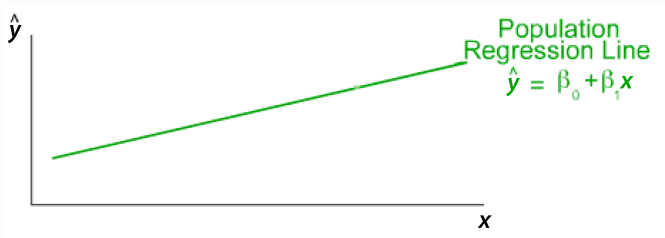
What happens in the population?

- Each of the sample regression coefficients, the y-intercept, b_0, and slope coefficient, b_1, has a corresponding population value
- **Notation**: The coefficients of the population regression line are β_0 and β_1, which define

  $$\hat{y} = \beta_0 + \beta_1 x$$

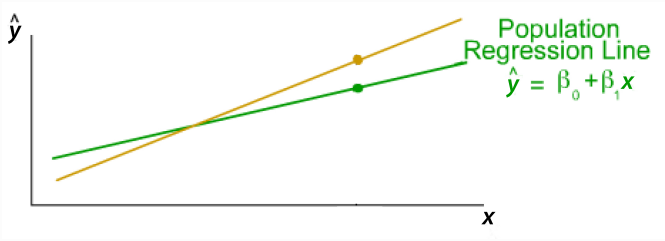
- For OLS regression estimates, unknown, but desired, population coefficients, β_0 and β_1, minimize $\sum e_i^2$ for the *entire* population

Population Regression Line



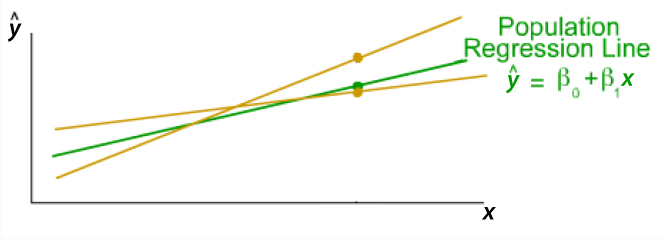
- The population coefficients, β_0 and β_1 describe the population regression line, which unfortunately cannot be directly observed

Population Regression Line



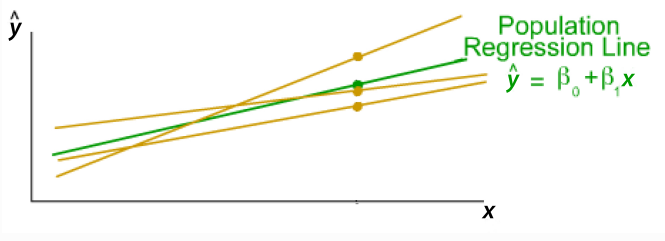
- Obtain a random sample and calculate the sample regression line, which differs from the population line because of sampling error

Population Regression Line



- Draw a second random sample and get another sample regression line because of sampling error, which leads to each sample regression line different from another

Population Regression Line



- Conceptually, the sample regression line randomly fluctuates from random sample to random sample even though only one sample is typically taken and only one regression line estimated

Inference of the Slope Coefficient: Focus on 0

What is the value of the population slope coefficient?

- The sample slope coefficients, b_1, the fitted values $\hat{y}_i$, and modeling errors or residuals, e_i, are all descriptive statistics
- Regarding the slope coefficient specifically: Is there a relationship *in the population* between the predictor variable, X, and the response variable y?
- For the *population* regression model,

$$\hat{y} = \beta_0 + \beta_1 x_1$$

- The focus of the inferential test for the population slope coefficient β_1 is on 0
  - If $\beta_1 < 0$, a negative relationship of X and y
  - If $\beta_1 = 0$, no relationship of X and y
  - If $\beta_1 > 0$, a positive relationship of X and y

Logic of Inference for the Slope Coefficient

What is the value of each population slope coefficient?

- Basic logic of confidence intervals and hypothesis tests for means and mean difference remains unchanged
- Here the statistic of interest is the sample slope coefficient, b_1, with corresponding population value β_1
- Inference is based on the t-cutoff, $t_{.025}$, and corresponding degrees of freedom, df
  - first pass through data calculates two values: b_0 and b_1
  - second pass through the *same* data to calculate residuals with the *two* previously estimated values, for b_0 and b_1
  - effective sample size, or degrees of freedom, is $df = n - 2$ for single predictor regression model

Hypothesis Test for the Slope Coefficient

How many estimated standard errors from the null?

- **Hypothesis Test for** β_1: Test the null hypothesis that $\beta_1 = 0$ with the t-statistic, the estimated number of standard errors the obtained b_1 is from 0, and then compare the resulting p-value to a pre-specified value of α, such as $\alpha = 0.05$
  - Null hypothesis: $H_0 : \beta_1 = 0$
  - Alternative hypothesis: $H_1 : \beta_1 \neq 0$
- If $p > .05$, conclude no difference detected from 0, that is, a relationship was not detected
- If $p < .05$, conclude a difference detected from 0, and, more informally, state the direction of the relation, positive or negative

Regression Output: Hypothesis Tests

Selectively edited output for the OLS estimates of the model

- The inferential analysis of a statistic begins with its estimated standard error, which, for the slope coefficient, is $s_{b_1} = 1.799$

	Estimate	Std Err	t-value	p-value
(Intercept)	-193.452	126.340	-1.531	0.164
Ht	5.202	1.799	2.892	0.020

- The corresponding t-value is $t_{b_1} = \dfrac{b_1 - 0}{s_{b_1}} = \dfrac{5.202}{1.799} = 2.892$
- **Statistical Decision**: p-value $= 0.020 < \alpha = 0.05$, so reject the null hypothesis of no relation of Heigth and Weight and, because $b_1 > 0$, conclude that $\beta_1 > 0$
- **Interpretation**: As Height increases, on average, Weight increases

Confidence Interval for the Slope Coefficient

What is the estimated value of the slope coefficient?

- **Confidence Interval** for the slope coefficient: Range of plausible values of β_1
- To construct the confidence interval, move $t_{.025}$ standard errors on either side of the sample slope coefficient

  β_1 within $b_1 \pm (t_{.025})(s_{b_1})$ for $df = n - 2$
- Is 0 is in the confidence interval?
  - If 0 is in the interval, then no relation detected between the corresponding predictor variable X and response variable y
  - If 0 is *not* in the interval, then all plausible values of the population slope coefficient, β_1, are either negative or positive, so either a negative or a positive relation has been detected between X and y

R: lessR reg Output, Confidence Intervals

Selectively edited output for the OLS estimates of the model

	Lower 95%	Upper 95%
(Intercept)	-484.794	97.889
Ht	1.054	9.349

- **Margin of Error**: $(t_{.025;n-2})(s_{b_1}) = (2.306)(1.799) = 4.15$
- The 95% confidence interval for β is

  from $5.202 - 4.15 = 1.05$ to $5.202 + 4.15 = 9.35$
- This confidence interval sets the likely bounds of the population slope coefficient, β, so apply the definition of β to each end of the interval
- **Interpretation**: At the 95% level of confidence, for each one inch increase in Height, on average, weight likely increases somewhere from slightly more than 1 pound to $9\frac{1}{3}$ pounds
- Margin of error is very large, but no surprise for just $n = 10$

Epilogue

Ultimate purpose is to interpret the statistical results

- The interpretation ...
  - generalizes the results from the sample to the population
  - without the use of jargon, explained in relaxed, conversational English
- The output by itself is worthless for business decision making
- The numbers on the output need to be translated into meaningful conclusions
- Use statistical jargon like "null hypothesis" and "p-value" for statistical reasoning, but no jargon for interpreting the results
- Statistical inference generalizes to the population
  - Hypothesis test tells us if a relationship exists, from which we can infer the direction, + or -
  - Confidence interval estimates the extent of the relationship, a range of plausible values of the population slope coefficient

- The End